Dynamic Extraction of Topic Descriptors and Discriminators: Towards Automatic Context-Based Topic Search

Ana Maguitman, David Leake, Thomas Reichherzer and Filippo Menczer
Computer Science Department, Indiana University
Lindley Hall 215, Bloomington, IN 47405, USA
{anmaguit, leake, treichhe, fil}@cs.indiana.edu

ABSTRACT

Effective knowledge management may require going beyond initial knowledge capture, to support decisions about how to extend previously-captured knowledge. Electronic *concept maps,* interlinked with other concept maps and multimedia resources, can provide rich *knowledge models* for human knowledge capture and sharing. This paper presents research on methods for supporting experts as they extend these knowledge models, by searching the Web for new context-relevant topics as candidates for inclusion. This topic search problem presents two challenges: First, how to formulate queries to seek topics that reflect the context of the current knowledge model, and, second, how to identify candidate topics with the right balance of novelty and relevance. More generally, this problem raises the broad question of the interaction of topic information from the local analysis space (a collected set of documents) and the global search space (the Web). The paper develops a framework for understanding this interaction, and proposes and evaluates techniques for addressing the query formation and topic identification questions by dynamically extracting topic descriptors and discriminators from a knowledge model, to characterize information needs for retrieval and filtering of relevant material. Using these techniques, we have developed a support tool that starts from a knowledge model under construction and automatically produces a set of suggestions for topics to include, proactively supporting users as they extend knowledge models.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation, Search process.*

General Terms

Algorithms, Experimentation.

Keywords

Knowledge Management, Concept Mapping, Context, Knowledge Acquisition Tools, Information Retrieval, Automatic Topic Search.

CIKM'04, November 8–13, 2004, Washington, DC, USA.

1. INTRODUCTION

An important question in knowledge management is how to determine the information to capture. In traditional views, knowledge capture may be seen as primarily knowledge acquisition, capturing knowledge that already exists within the expert. This paper presents methods for supporting an alternative approach, "knowledge extension," based on the premise that a knowledge model evolves from coordinated processes of knowledge acquisition and knowledge construction. In this view, it is crucial to support experts' construction of new knowledge as they extend existing knowledge models. The paper addresses this need by first developing the theoretical framework required for a "topic suggester" to propose candidate topics—possible themes—for new concept maps to add to a knowledge model, and then evaluating an implemented system based on that framework.

The World Wide Web provides a rich source of information on potential new topics to include in a knowledge model. To access relevant information, appropriate queries must be formed. In text-based Web search, users' information needs and candidate text resources are typically characterized by terms. Substantial experimental evidence supports the effectiveness of using weights to reflect relative term importance for traditional information retrieval (IR) [Salton and Yang, 1973, Salton and Buckley, 1988]. However, as has been discussed by a number of sources, issues arise when attempting to apply conventional IR schemes for measuring term importance to systems for searching Web data [Kobayashi and Takeda, 2000, Belkin, 2000]. One difficulty is that methods for automatic query formation for Web search do not have access to a full predefined collection of documents, raising questions about the suitability of classical IR schemes for measuring term importance when searching the Web. In addition, the importance of a given term depends on the task at hand; the notion of term importance has different nuances depending on whether the term is needed for query construction, index generation, document summarization or similarity assessment. For example, a term which is a useful descriptor for the content of a document, and therefore useful in similarity judgments, may lack discriminating power, rendering it ineffective as a query term, due to low precision of search results, unless it is combined with other terms which can discriminate between good and bad results. The central question addressed in this paper is how to formulate topic descriptors and discriminators to guide this search process.

The IR community has investigated the roles of terms as descriptors and discriminators for several decades. Since Sparck Jones' seminal work on the statistical interpretation of term specificity [Jones, 1972], term discriminating power has often been interpreted statistically, as a function of term use. Similarly, the importance of terms as content descriptors has been traditionally estimated by measuring the frequency of a term in a document. The combination of descriptors and discriminators gives rise to schemes for measuring term relevance such as the familiar *term frequency inverse document frequency* (TFIDF) weighting model [Salton and Yang, 1973]. The task domain for our research is support for the extension of concept-map-based *knowledge models*, by proposing new topics to include in a collection of concept maps [Leake et al., 2003a, Leake et al., 2003b]. Searching the Web to support this knowledge extension process presents new challenges for formulation of descriptors and discriminators. Specifically, making full use of the information available in these knowledge models requires:

- **Search methods that can reflect extensive contextual information** (instead of attempting to summarize context in a small number of terms). For knowledge model extension, the knowledge model under construction provides a rich context that can be exploited for information filtering, term-weight reinforcement, and query refinement. Because search engines may restrict queries to a small number of terms (e.g., the 10-term limit for Google), incremental approaches may be needed to fully reflect search context.

- **Methods for topic search** (instead of document search). Users selecting topics to include in a knowledge model will be aided by search methods which directly generate characterizations of possible topics—which may span individual documents—rather than simply presenting sets of documents. In traditional IR approaches, term discriminating power is based on the overall rarity of a term in a document collection, rather than on term distribution across different topics. For example, the term discrimination value under the TFIDF model expresses the goodness of a term in discriminating a *document*, as opposed to discriminating the *topic* of the document. Mining topics requires new measures for term discrimination.

- **Methods for searching open collections of documents** (instead of a pre-defined and pre-analyzed collection). In Web-based knowledge extension tasks, the search space is the full Web, and analysis must be limited to a small collection of documents—incremental retrievals—that is built up over time and changes dynamically. Unlike traditional IR schemes, which analyze a predefined collection of documents and search that collection, Web-based knowledge extension must rely on methods that use limited information to assess the importance of documents and to manage decisions about which documents to retain for further analysis, which ones to discard, and which additional queries to generate.

This paper introduces and evaluates techniques that automatically formulate queries from knowledge models, and analyze the results returned by a Web search engine to dynamically discover topic descriptors and discriminators for topics related to the initial knowledge model. Unlike standard techniques for the discovery of descriptors and discriminators, the approach does not estimate term distributions across a predefined collection of documents. Instead, it uses an incrementally-retrieved, topic-dependent selection of documents for term-weight reinforcement reflecting the aptness of the terms in describing and discriminating the topic in question. The methods find new descriptors by searching for terms that tend to occur *often* in relevant documents, and find good discriminators by identifying terms that tend to occur *only* in the context of the given topic. The methods have been tested using the Google Web API service. However, they could also be used to query other sources, either from commercial services or—if those services became expensive or unavailable—alternative sources such as indices generated by topical crawlers (e.g., [Chakrabarti et al., 1999, Menczer et al., 2004]).

2. AIDING KNOWLEDGE EXTENSION WITH TOPIC SUGGESTIONS

The practical motivation for our work is the development of a topic suggester tool for aiding extension of knowledge models based on electronic concept maps. Concept mapping has been widely used as a vehicle for knowledge capture and sharing, both in educational and commercial settings (see [Cañas et al., 1995, Ford et al., 1996, Cañas et al., 1998, Hoffman et al., 2001] for a sampling of work in this area). Concept mapping, developed by Novak for use in education, is designed as a vehicle for making cognitive structures explicit by externalizing the concepts and propositions known to a person [Novak and Gowin, 1984], but the process of concept mapping is also viewed as a means to aid people in constructing meaningful knowledge, by organizing their knowledge and making relationships explicit.

Concept maps are two-dimensional graphical representations of a set of concepts, connected by directed arcs encoding propositions in the form of simplified sentences, to show their interrelationships. The vertical layout tends to express a hierarchical framework for the concepts, with inclusive concepts usually found at the highest levels and progressively more specific concepts arranged below them. Portions of two concept maps about the topic of Mars exploration are shown in Figure 1. It is important to note that concept maps are a vehicle for *human* knowledge sharing, rather than a formal framework intended for automated reasoning.

CmapTools, developed by the Institute for Human and Machine Cognition (IHMC), is a suite of publicly-available software tools for knowledge acquisition, construction, and sharing (http://cmap.ihmc.us) based on concept maps. The CmapTools system provides an easy-to-use interface for human knowledge capture, extension, and examination. It has been used by a wide range of people, from elementary school children to NASA scientists. The software enables experts to construct knowledge models of their domain without the need for a knowledge engineer's intervention, or to actively participate in the knowledge elicitation if a knowledge engineer leads the process.

The CmapTools effort includes a collaboration between researchers at IHMC and Indiana University to develop tools to aid the concept mapping process. The tools are designed to address difficulties which have been observed arising during concept mapping. For example, users sometimes stop and wonder what concepts to add to a concept map; spend time trying to find the right word to use in a concept label or linking phrase; search for relevant concept maps to compare; and search the Web for additional material to enhance the concept map or to jog their memories for topics to include. Each of these has been addressed by a system to suggest relevant information, based on the context provided by the concept map. Each system starts from a concept map under construction, and proactively suggests relevant information such as concept maps, propositions, resources, concepts and topics. The suggesters are described in detail in [Leake et al., 2003b].

This paper focuses on issues and algorithms for EXTENDER, CmapTools' topic suggester. Starting from a concept map, EX-

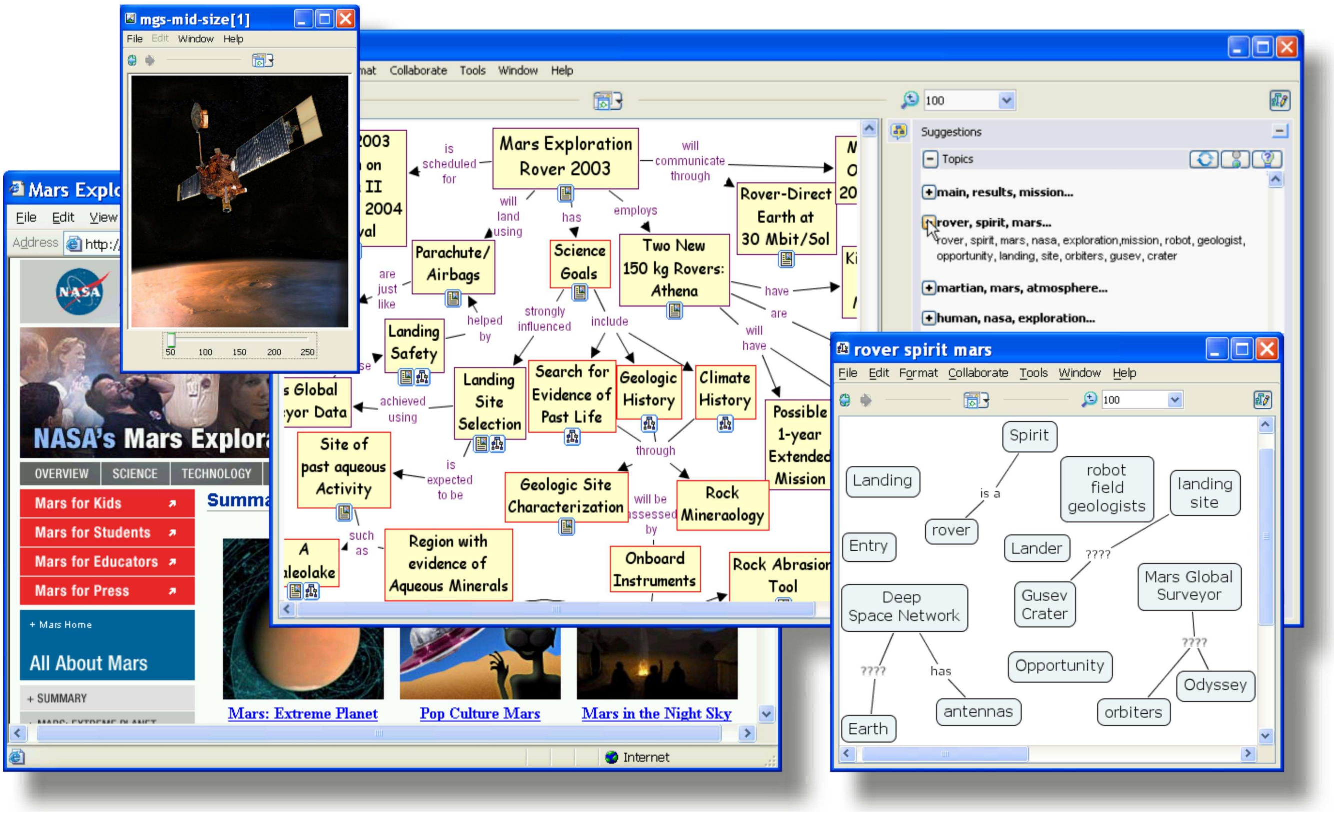


Figure 1: Portion of a Knowledge Model with EXTENDER suggesting new topics.

TENDER identifies and suggests sets of terms characterizing novel but related topics, as candidate new topics for inclusion in a knowledge model. In addition, it organizes the Web pages that gave rise to those topics according to topic, to facilitate access to topic-relevant information. Figure 1 shows part of a knowledge model with EXTENDER's suggestion window for new topics at the upper right. The in-progress concept map in the bottom right contains some concepts that the user selected from a topic suggested by EXTENDER.

Given a concept map, EXTENDER extracts terms that serve as topic descriptors and discriminators, and uses those terms to guide Web search for sets of terms corresponding to related but novel topics. A key goal is for these term sets to exhibit both *global coherence*—the suggested topics must be relevant to the topic of the originating concept map—and *local coherence*—the terms selected for inclusion in the generated topic set must be cohesive with each other. Both topic descriptors and topic discriminators play fundamental roles in our approach to achieving *global coherence*, the focus of this paper. A soft clustering algorithm tailored for EXTENDER, not discussed here, addresses the requirement of attaining local coherence.

3. A FRAMEWORK FOR ANALYZING DOCUMENTS, TERMS AND TOPICS

Topics group documents related by a common theme. One way to represent topics is implicitly, as sets of related documents. Alternatively, a topic can be represented as a set of cohesive terms summarizing the topic content. Some terms may have strong descriptive power, enabling a small set to convey the topic to a human. Other terms may be effective cues for retrieving topic-relevant documents, but may not be good descriptors. Consider for example a topic involving exploration of Mars, described by the following set of terms occurring in documents related to Mars exploration:

Mars	Exploration	Rover	Landing	Site
Selection	Opportunity	Spirit	Images	Global
Surveyor	Orbiter	Camera	MGS	MOC

The terms *Mars* and *Exploration* are good descriptors of the topic for a general audience. Terms such as *MGS* and *MOC*—which stand for "Mars Global Surveyor" and "Mars Orbiter Camera"—may not be good descriptors of the topic for that audience, but are effective in bringing information similar to the topic when presented in a query.

Intuitively, we can characterize topic descriptors and discriminators as follows:

- Terms are *good topic descriptors* if they answer the question "What is this topic about?"
- Terms are *good topic discriminators* if they answer the question "What are good query terms to access similar information?"

Our hypothesis, examined in this paper, is that terms that tend to occur frequently in the context of a given topic tend to be good topic descriptors. Thus a possible strategy for finding good topic descriptors is to (1) find documents that are similar to other documents already known to have that topic, and (2) select from those documents the terms that occur often. On the other hand, a term is a good discriminator for a topic if most documents that contain that term are topically related. Thus finding good topic discriminators requires finding terms that tend to occur only in the context of the given topic. Both topic descriptors and discriminators are important as query terms. Because topic descriptors occur often in relevant pages, using them as query terms may improve recall. Because good topic discriminators occur primarily in relevant pages, using discriminators as query terms may improve precision. The following sections transform the above informal characterizations of topic descriptors and discriminators into precise definitions and apply them to the task of searching the Web for context-related topics.

3.1 Using Hypergraph Representations for Documents and Terms

Determining topic discriminators and descriptors requires analyzing the interplay between terms, documents and topics. We propose hypergraphs [Berge, 1973] as a natural way to represent such relationships. A hypergraph is a generalization of a graph, in which each edge (hyperedge) is represented as a multiset of nodes. If we disregard the structure of text documents, we can view any collection of documents as a hypergraph $H = (T, \mathcal{D})$, where each node $t \in T$ corresponds to a term and each hyperedge $d \in \mathcal{D}$ corresponds to a document. A hyperedge d is a multiset with elements in T, representing the abstraction of a document as a bag of terms. We call this a *document-centered hypergraph*. As a dual to this view, we can think of a term as a multiset whose elements are those documents in which the term occurs. Therefore, for each document-centered hypergraph $H = (T, \mathcal{D})$, there corresponds a *term-centered hypergraph* $H^* = (D, \mathcal{T})$ whose nodes correspond to documents and whose hyperedges correspond to terms, represented as multisets of documents. Hypergraph H^* is called the dual hypergraph of H. Figures 2(a) and 2(b) illustrate a hypergraph representation for a collection of three documents, A, B, and C, each represented as a multiset, containing some of the terms 1, 2, 3 and 4. This collection can be represented by the document-centered hypergraph $H = (\{1, 2, 3, 4\}, \{A, B, C\})$ (with $A = \{1, 1, 2\}, B = \{2, 2\}$ and $C = \{2, 3, 4\}$) or by its dual $H^* = (\{A, B, C\}, \{1, 2, 3, 4\})$ (with $1 = \{A, A\}, 2 = \{A, B, B, C\}, 3 = \{C\}$ and $4 = \{C\}$.) In figures 2(a) and 2(b), circles represent hyperedges and triangles represent nodes. The value associated with the connection between a node and a hyperedge stands for the number of occurrences of the node in the hyperedge. For example, the value 2 associated with the connection between node 1 and hyperedge A in figure 2(a) denotes that term 1 occurs twice in document A.

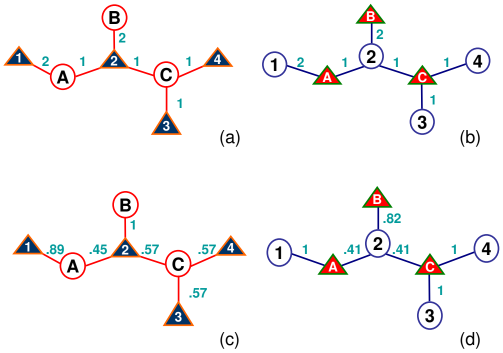


Figure 2: (a) hypergraph H; (b) hypergraph H^*; (c) and (d) the hypergraphs' weighted version.

The incidence matrix of a document-centered hypergraph $H = (T, \mathcal{D})$ for a collection of m documents and n terms is a matrix $\mathbf{H}$ with m rows that represent the documents (hyperedges of H) and n columns corresponding to the terms (nodes of H) such that

$$\mathbf{H}[i, j] = k$$

where k is the number of occurrences of t_j in d_i. Note that the incidence matrix of the dual hypergraph H^* is the transpose of the incidence matrix of hypergraph H.

Representing the relationships between terms and documents using hypergraphs forms the basis for our analysis of a series of dual notions. These dualities arise at various levels, and can be interpreted as reflecting interesting properties of terms and documents leading to our characterization of topic descriptors and discriminators.

3.2 Document Descriptors and Discriminators

We use the adjacency matrix $\mathbf{H}$ of a document-centered hypergraph to define functions corresponding to the notions of term descriptive power and term discriminating power in a document. Term descriptive power in a document is modeled by a function $\lambda : \{d_0, \ldots, d_{m-1}\} \times \{t_0, \ldots, t_{n-1}\} \rightarrow [0, 1]$ that maps a document-term pair into a value in the unit interval. It is defined as follows:

$$\lambda(d_i, t_j) = \frac{\mathbf{H}[i, j]}{\sqrt{\sum_{k=0}^{n-1} (\mathbf{H}[i, k])^2}}.$$

Function λ can be used to construct a *document-centered weighted hypergraph* (which we will call a d-hypergraph) in which the descriptive power of term t_j in document d_i is used as the weight of node t_j in hyperedge d_i. In figure 2(c) we can see a d-hypergraph in which terms have different descriptive power for their associated documents. In particular, document B is entirely described by term 2.

The second function $\delta : \{t_0, \ldots, t_{n-1}\} \times \{d_0, \ldots, d_{m-1}\} \rightarrow [0, 1]$ is used to model discriminating power of a term in a document. If we define $\mathrm{s}(k)$, to return 1 if $k > 0$ and 0 if $k = 0$, we define δ as follows:

$$\delta(t_i, d_j) = \frac{\mathrm{s}(\mathbf{H}^{\mathbf{T}}[i, j])}{\sqrt{\sum_{k=0}^{m-1} \mathrm{s}(\mathbf{H}^{\mathbf{T}}[i, k])}}.$$

Function δ maps a term-document pair into a value in the unit interval. If term t_i does not occur in document d_j then $\delta(t_i, d_j) = 0$. On the other extreme, if term t_i occurs in no document other than d_j, then $\delta(t_i, d_j) = 1$ and we say that t_i fully discriminates d_j. Discriminating power of a term in a document is independent of the number of occurrences of the term in the document. If k represents the number of occurrences of a term in a document, function δ will only consider $\mathrm{s}(k)$, disregarding the total number of occurrences and considering only whether or not a term is in a document. Function δ can be used to construct a *term-centered weighted hypergraph* (t-hypergraph) where the discriminating power of term t_i in document d_j is the weight of node d_j in hyperedge t_i. In figure 2(d), term 1 fully discriminates document A.

Both for d-hypergraphs and t-hypergraphs, the square of the weights associated with each hyperedge sum to 1, i.e.,

$$\sum_j (\lambda(d_i, t_j))^2 = 1 \quad \text{and} \quad \sum_j (\delta(t_i, d_j))^2 = 1.$$

It is easy to verify that the weighted hypergraphs will continue to be duals structurally, but in general they will not preserve the numerical duality. Consequently, the new associated incidence matrices will not be transposes of each other.

As is the case with other IR characterizations of descriptors and discriminators, the notions discussed above only allow discovering terms that are good descriptors or discriminators of a *document*, as opposed to good descriptors or discriminators of the *topic of a document*. In the next sections, we build on the notions of document descriptors and discriminators to identify higher-order relationships between documents and terms and to provide new definitions of descriptors and discriminators. These new definitions make the notions of descriptors and discriminators topic-dependent.

3.3 Similarity and Co-occurrence

To address the problem of identifying terms that are good descriptors or good discriminators of a topic, we first need to characterize the notion of *topic*. We treat topics as defined by either a collection of similar documents or a collection of terms that tend to co-occur. Thus the notions of document similarity and term co-occurrence play important roles in identifying topics.

The similarity between documents d_i and d_j can be computed using the well-known cosine measure as follows:

$$\begin{aligned}\sigma(d_i, d_j) &= \frac{\sum_{k=0}^{n-1}(\lambda(d_i,t_k)\cdot\lambda(d_j,t_k))}{\sqrt{\sum_{k=0}^{n-1}(\lambda(d_i,t_k))^2\cdot\sum_{k=0}^{n-1}(\lambda(d_j,t_k))^2}}\\ &= \sum_{k=0}^{n-1}(\lambda(d_i, t_k)\cdot\lambda(d_j, t_k)).\end{aligned}$$

The idea of *term co-occurrence* captures a relation between terms that is dual to the notion of document similarity. If two terms tend to occur in the same documents, it is likely that their meanings are related. A measure of co-occurrence for terms t_i and t_j can be obtained as follows:

$$\begin{aligned}\kappa(t_i, t_j) &= \frac{\sum_{k=0}^{m-1}(\delta(t_i,d_k)\cdot\delta(t_j,d_k))}{\sqrt{\sum_{k=0}^{m-1}(\delta(t_i,d_k))^2\cdot\sum_{k=0}^{m-1}(\delta(t_j,d_k))^2}}\\ &= \sum_{k=0}^{m-1}(\delta(t_i, d_k)\cdot\delta(t_j, d_k)).\end{aligned}$$

Figure 3(a) presents a simple illustration of the notion of document similarity by means of a d-hypergraph. In this example we can see that documents D and E are similar (they share the terms *mars*, *missions* and *nasa*.) Figure 3(b) shows the corresponding t-hypergraph in which it is easy to see that terms 3 (*missions*) and 4 (*nasa*) co-occur.

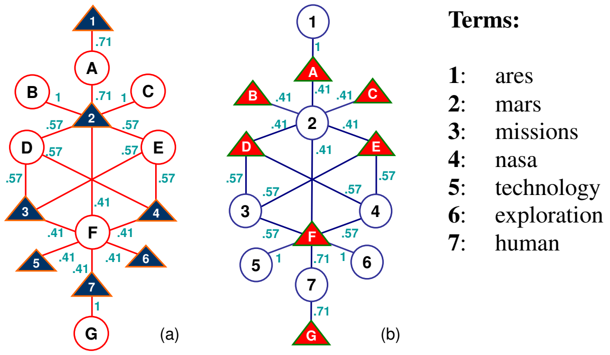


Figure 3: Weighted hypergraphs illustrating a series of dual notions: document similarity, term co-occurrence, topic discriminators, topic focus, topic descriptors and topic exhaustivity.

3.4 Topic Discriminators and Topic Focus

By examining document-term duality, we can develop higher-order notions useful for identifying good topic descriptors and discriminators. A term is a *good discriminator of a document's topic* if those documents discriminated by the term are similar to the given document. This intuition can be formally expressed using the function $\Delta : \{t_0, \ldots, t_{n-1}\} \times \{d_0, \ldots, d_{m-1}\} \rightarrow [0, 1]$ defined as follows:

$$\Delta(t_i, d_j) = \sum_{\substack{k=0\\k\neq j}}^{m-1}(\delta(t_i, d_k)^2\cdot\sigma(d_k, d_j)).$$

We can think of the discriminating power of term t_i for the topic of document d_j as the average of the similarity of d_j to other documents discriminated by t_i. Note that even in the case when d_j does not contain t_i, the value of the function $\Delta(t_i, d_j)$ will not necessarily be 0. On the other hand, if no other document similar to d_j contains t_i, i.e., $\sigma(d_k, d_j) = 0$ or $\delta(t_i, d_k) = 0$ for all documents d_k containing t_i with $k \neq j$, then t_i has no discriminating power over the topic of d_j and as a consequence $\Delta(t_i, d_j) = 0$.

We have previously discussed the dual notions of document similarity and term co-occurrence. At this stage we might ask what would be the dual notion to "term discriminating power in a topic." This would be a function comparable to Δ but applicable to documents rather than terms. We can think of *document focus* as a property of documents that plays a role dual to that of *term discriminating power*. A document is focused on the topics associated with a term if the terms describing the document tend to co-occur with the given term. Formally, we can compute the degree of focus of a document on the topic identified by a term as a function $\Phi : \{d_0, \ldots, d_{m-1}\} \times \{t_0, \ldots, t_{n-1}\} \rightarrow [0, 1]$ defined as follows:

$$\Phi(d_i, t_j) = \sum_{\substack{k=0\\k\neq j}}^{n-1}(\lambda(d_i, t_k)^2\cdot\kappa(t_k, t_j)).$$

Note that we have defined the higher-order dual notions of topic discriminators and topic focus by means of more basic dual notions. Term discriminating power in a topic has been defined using the notions of term discriminating power in a document and document similarity. Analogously, the measure of document focus on a topic has been defined via term descriptive power in a document and term co-occurrence.

3.5 Topic Descriptors and Topic Exhaustivity

The notion of *topic descriptors* was informally defined earlier as terms that occur *often* in the context of a topic. The *descriptive power* of a term in a topic is a measure that can be computed using the previously defined measures of document similarity and term descriptive power in documents. We measure *term descriptive power in the topic of a document* as a function $\Lambda : \{d_0, \ldots, d_{m-1}\} \times \{t_0, \ldots, t_{n-1}\} \rightarrow [0, 1]$:

$$\Lambda(d_i, t_j) = \begin{cases} 0 & \text{if } \sum_{\substack{k=0\\k\neq i}}^{m-1}\sigma(d_i, d_k) = 0\\ \dfrac{\sum_{\substack{k=0\\k\neq i}}^{m-1}(\sigma(d_i,d_k)\cdot\lambda(d_k,t_j)^2)}{\sum_{\substack{k=0\\k\neq i}}^{m-1}\sigma(d_i,d_k)} & \text{otherwise.}\end{cases}$$

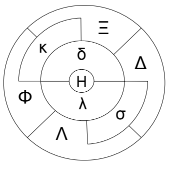
Descriptive power of a term t_j in the topic of a document d_i is a measure of the quality of t_j as a descriptor of documents similar to d_i. If no other document is similar to d_i or t_j does not occur in other documents similar to d_i then the descriptive power of t_j in the topic of d_i is equal to 0.

The last property we define is *document exhaustivity* with regard to a topic. A document is exhaustive (or comprehensive) with regard to the topic identified by a term if most terms that co-occur with the given term tend to discriminate that document; exhaustivity of a document can be thought of as the dual property of descriptive power of a term. We propose a measure of document exhaustivity as a function $\Xi : \{t_0, \ldots, t_{n-1}\} \times \{d_0, \ldots, d_{m-1}\} \rightarrow [0, 1]$:

$$\Xi(t_i, d_j) = \begin{cases} 0 & \text{if } \sum_{\substack{k=0\\k\neq i}}^{n-1} \kappa(t_i, t_k) = 0 \\ \dfrac{\sum_{\substack{k=0\\k\neq i}}^{n-1} (\kappa(t_i,t_k) \cdot \delta(t_k,d_j)^2)}{\sum_{\substack{k=0\\k\neq i}}^{n-1} \kappa(t_i,t_k)} & \text{otherwise.} \end{cases}$$

By the definition of $\Xi(t_i, d_j)$, if term t_i does not co-occur with any other term or d_j does not contain any term that co-occurs with t_i then the exhaustivity of d_j with regard to the topic associated with t_i is 0.

In the hypergraphs of figure 3 terms 2 (*mars*), 3 (*missions*) and 4 (*nasa*) are all good descriptors in the topic of documents D, E and F. However, while terms 3 and 4 are good discriminators in that topic, term 2 is not—term 2 occurs often in that topic but not only in that topic. Note also that in this example documents D, E and F are exhaustive on the topic of terms 2, 3, and 4. Among these three documents, only D and E are focused on the topic. For example, document F contains most terms that co-occur in that topic but not only terms from that topic. The diagram of figure 4 summarizes the notions discussed in this section. It starts with the hypergraph incidence matrix $\mathbf{H}$ in the center of the diagram, where $\mathbf{H}[i, j]$ represents the *number of occurrences of term* t_j *in document* d_i, and shows how the higher-level notions are built upon the more basic ones. Dual notions (e.g., similarity and co-occurrence) appear on opposite sides of the diagram.



$\mathbf{H}[i, j]$: number of occurrences of t_j in d_i
$\delta(t_i, d_j)$: discriminating power of t_i in d_j
$\lambda(d_i, t_j)$: descriptive power of t_j in d_i;
$\sigma(d_i, d_j)$: similarity of documents d_i and d_j.
$\kappa(t_i, t_j)$: co-occurrence of terms t_i and t_j.
$\Delta(t_i, d_j)$: discriminating power of t_i in d_j's topic.
$\Lambda(d_i, t_j)$: descriptive power of t_j in d_j's topic.
$\Phi(d_i, t_j)$: focus of d_i on t_j's topic.
$\Xi(t_i, d_j)$: exhaustivity of d_j on t_i's topic.

Figure 4: The different levels of the document-term duality

The higher-order notions of discriminating power, descriptive power, focus and exhaustivity are useful for identifying and characterizing topics. Topic descriptors and discriminators are useful as query terms to favor recall and precision respectively. We have applied discriminating power and focus in the implementation of a clustering algorithm (not discussed in this paper) to produce cohesive topics. Because descriptors describe the subject of a topic, they are good terms to use as the topic's label, when the topic is presented to the user. A combination of focus and exhaustivity can be used to rank documents in a topic.

The rest of this paper addresses the issue of extracting good topic descriptors and good topic discriminators. The notions of focus and exhaustivity have been introduced for completeness and will be examined in a separate paper.

4. APPLYING THE THEORY

The framework developed in the previous section has been applied in the implementation of EXTENDER. EXTENDER starts from a concept map and iteratively mines the Web, searching for novel information, which is clustered to produce topics that are related to the initial concept map. At each iteration the system's goal is to extend the current topics, an operation that requires searching the Web for related novel material. Because retrieving and processing large numbers of Web pages is costly, EXTENDER first applies a less expensive *distillation phase*, in which a series of queries is submitted to a search engine and only the information that is readily available from the search results (e.g. title, "snippet" of text, url, Open Directory Project summary) is used to identify good topic descriptors and discriminators. After this preliminary step, the best topic descriptors and discriminators are used as query terms in a *search phase* to search for additional material on the Web. The collected material is represented by means of hypergraphs, unimportant terms are discarded, and clustering is applied to identify topics in the collection. The clustering phase is implemented by a hypergraph-based soft clustering algorithm (not discussed here) tailored for EXTENDER. This process is repeated a number of times, with the stopping criterion depending on settings selected by the user.

EXTENDER uses a "curiosity mechanism" to favor exploration during initial processing stages and exploitation towards the end. Throughout the exploration phase, while attempting to extend a given topic $\mathbf{T}$, new-found terms are collected. For each term t, the system tracks both the goodness of t in describing the topic $\mathbf{T}$ and the goodness of t in discriminating $\mathbf{T}$. To do so, it considers $\mathbf{T}$ as a multiset of terms and computes functions $\Lambda(\mathbf{T}, t)$ and $\Delta(t, \mathbf{T})$, respectively. Because the number of collected terms grows rapidly, novel terms are only preserved if they survive a selection process. For iteration $\textsc{i}$, the threshold for the survival of descriptors is computed by means of a function $\mathcal{T}_\Lambda : \{0, \ldots, s-1\} \rightarrow [a, b]$:

$$\mathcal{T}_\Lambda(\textsc{i}) = (b - a) \cdot \left(\frac{\textsc{i}}{s-1}\right)^c + a,$$

where a stands for the "least threshold" parameter, b for the "greatest threshold" parameter, c is a curiosity decay parameter, and s is the total number of iterations. The parameter a (resp. b) reflects the initial (final) stage of exploration (exploitation), when many (few) new terms are collected. The threshold for discriminators, $\mathcal{T}_\Delta$, is defined similarly. Because the curiosity threshold increases with the number of iterations, novel terms are seldom collected during the final stages. As a consequence, the exploitation phase primarily reinforces the weights associated with particular terms that have been already added to the collection. Another curiosity threshold is used by EXTENDER to filter irrelevant documents. This is implemented by a similarity threshold function $\mathcal{T}_\sigma$ defined analogously to the definition of function $\mathcal{T}_\Lambda$. Figure 5 presents a high-level description of the topic extension algorithm.

5. EXPERIMENTAL STUDY

It is relatively simple to evaluate the effectiveness of techniques for selecting good discriminators to use as query terms. This can be done by providing an approximate measure of the relevance of the retrieved documents (e.g., by measuring the mean similarity between the retrieved documents and the source) and using that relevance measure to compare the performance of the new technique against baseline techniques. In section 5.2 we report a controlled study to evaluate the distillation method for query formation proposed in this paper. However, it is more difficult to develop objective measures for evaluating descriptive power. In this study, we propose the use of concept map libraries as data for assessing term descriptive power. From a data-processing perspective, concept maps present an important advantage over purely textual forms in at least two respects: (1) in concept maps, concepts and their relationships are readily available, and (2) concept maps are

```
ALGORITHM
INPUT:
  T: input topic (or source concept map);
  $s$: total number of iterations;
  $q_d$: number of queries submitted for distillation;
  $q_s$: number of queries submitted for search;
OUTPUT:
  Topics: A set of topics related to T.
BEGIN
  Topics[0]:= {T}
  FOR ($i := 0$; $i < s$; i++) DO
    Topics[i+1]:=∅.
    FOR EVERY Topic ∈ Topics[i] DO
      $N$ := NextGenerationOfTopics(Topic, $i$).
      Topics[i+1]:= Topics[i+1] ∪ N.
    END DO
    Merge similar topics in Topics.
  END DO
  RETURN Topics.
END

PROCEDURE NextGenerationOfTopics
INPUT:
  T: topic to extend;
  $i$: present iteration;
OUTPUT:
  $N$: A new set of topics.
BEGIN
  //distillation
    Use the terms $t$ with highest $\lambda(\mathbf{T}, t)$ value to form $q_d$ queries.
    Submit the queries to a search engine.
    Only keep documents $d$ such that $\sigma(d, \mathbf{T}) \geq \mathcal{T}_\sigma(i)$.
    Use search result's "readily available information" to compute
      $\Lambda(\mathbf{T}, t)$ and $\Delta(t, \mathbf{T})$ for each term $t$.
  //search
    Combine the terms $t$ with highest $\Delta(t, \mathbf{T})$ value and the terms with
      highest $\Lambda(\mathbf{T}, t)$ value to form $q_s$ queries.
    Submit the queries to a search engine.
  //filtering
    Only collect documents $d$ such that $\sigma(d, \mathbf{T}) \geq \mathcal{T}_\sigma(i)$.
    Only keep terms $t$ such that $\Delta(t, \mathbf{T}) \geq \mathcal{T}_\Delta$(i) or $\Lambda(\mathbf{T}, t) \geq \mathcal{T}_\Lambda$(i).
  //clustering
    Cluster collected data to generate a set $N$ with new topics.
  RETURN $N$.
END
```

Figure 5: Pseudocode of the Topic Extension Algorithm

usually hierarchical and have a rich topology. Our previous studies show that topological analysis algorithms can be adapted to the analysis of concept maps to describe the relative arrangements of their concepts, and that the topological roles of concepts in the map can be usefully summarized according to a small set of dimensions [Cañas et al., 2001]. Our previous studies also provide evidence for the significance of topological factors in human assessments of concept descriptive power in concept maps [Leake et al., 2004]. In section 5.1 we first summarize a previous model in which topological factors are used to assess concept descriptive power in concept maps, and its fit with human-subjects data. Then, we make use of this model to *indirectly* evaluate the prediction power of the topic-descriptor extraction algorithm.

5.1 Evaluating the Descriptor Extraction Method

5.1.1 Modeling Concept Descriptive Power

We previously developed three candidate models of the importance of concepts in describing the content of concept maps, and evaluated their fit with data on human judgments [Leake et al., 2004]. These models use the topology of concept maps to compute a weight predicting each concept's importance in describing the topic of a map. To determine which factors to include in the models, we first considered factors from the concept mapping literature. For example, Novak proposes that concept maps should have a hierarchical structure. The candidate models can reflect such a structure, with weightings reflecting that more descriptive concepts are at the top of the map, and less descriptive at the bottom.

As a starting point for evaluating the descriptor extraction method, we consider the *path frequency* model (PF), which reflects the expectation that concepts participating in more propositions will tend to be more important as descriptors of the topic of a map. The *path frequency* measure can be seen as the concept-map counterpart of the *term frequency* measure in the TFIDF scheme. (Term frequency is not a good estimator of term descriptive power in a concept map, because each term rarely occurs more than once in a concept map). The PF model counts all possible paths, starting from the root concept, that contain the concept in question and either (1) end on a concept with no outgoing connections, or (2) end on a concept that has already been visited. The weight $W_{PF}(c)$ of a concept c in a map is the number of paths crossing c. We note that if a concept has high connectivity (which allows for many paths to form in the map), then the number of paths crossing a concept also increases for concepts indirectly linked to the high-connectivity concept. Due to the hierarchical structure of concept maps, concepts that are closer to the root tend to participate in more paths. In particular, the root concept participates in all possible paths in a map and as a consequence it receives the highest PF weight.

We conducted a human-subjects experiment to study the influences of the hypothesized factors on human judgments of concept descriptive power, and the overall fit of our models' predictions to human judgments. Twenty paid subjects, all students admitted to Indiana University, participated in the study. Subjects answered 56 questions about a total of 12 small concept maps (fewer than 15 concepts each). Two of the concept maps were used in a training phase while the remaining 10 were used for the test. Each question presented a concept map and two concepts selected from that map. Participants were asked to examine the map and to answer which of the two concepts best described the map's topic, or whether both described it equally well. The root-mean-square error (RMSE) between the user and PF model data was of 0.170. Details on the study method and results can be found in [Leake et al., 2004].

The results show that the PF model provides a good fit to the user data, suggesting that it can be used as a target to indirectly evaluate methods aimed at assessing the descriptive power of terms in a topic—provided we have access to a concept map representation of the topic as a starting point.

5.1.2 Using the Topology Analysis Model to Evaluate Description Extraction

We took advantage of the fit of the PF model to human data to perform an indirect evaluation of the descriptor extraction method by means of concept maps. As data we used the Mars 2001 knowledge model, a large multimedia knowledge model on Mars (http://www.cmex.arc.nasa.gov), constructed entirely by NASA scientists using CmapTools. The Mars 2001 knowledge model contains 118 concept maps and 3654 concepts. Our goal in this evaluation was to test if the descriptor extraction method discussed in this paper was able to predict the weights assigned by the PF model.

We used each concept in a concept map to submit a query to GOOGLE (using the GOOGLE Web API) and up to 20 results were collected for each query (approximately 600 Web pages were collected for each concept map). The queries were constructed using all the terms in a concept label, after stop-word filtering and disregarding the topological role of the concept in the map. For example, the concept "Search for evidence of Past Life" from the map of figure 1, was presented to GOOGLE as '*search* AND *evidence* AND *past* AND *life*'. For each concept map **M** in the Mars 2001 project we tested if the descriptor-extraction method was able to predict the topological term weighting suggested by the PF model. In order to do so, given a concept map **M** and a collection of retrieved Web pages, we computed the $\Lambda(\mathbf{M}, t)$ measure for each term in the collection. Results were compared to a baseline model in which all terms in a map were assigned the same weight.

The RMSE between the PF model data and the descriptor-extraction method (Λ) was of 0.237 while the RMSE between the PF model and the baseline model was 0.824. Table 1 summarizes the RMSE for each test. In addition, the Pearson correlation coefficient between the PF model weighting and that of the descriptor-extraction method was 0.42 for 6901 pairs, where the pairs contain the PF and Λ weights of the terms found in the Mars 2001 knowledge model. This result indicates a statistically significant correspondence between the two weighting schemes. Hence, by transitivity, the combination of this result with the results obtained in the previously reported human subject experiment suggests a considerable correspondence between human judgments of concept descriptive power and the data returned by the descriptor-extraction method. This correspondence is encouraging for the hypothesis that the proposed method provides good predictions on the importance of terms in describing a topic.

	USER DATA	Λ	BASELINE
PF	0.170	0.237	0.824

Table 1: Summary of RMSE of PF compared to user data, Λ, and baseline.

As a sidenote, it is interesting to note that the Pearson correlation coefficient between the PF model weighting and that of the discriminator-extraction method was only 0.01. This result reflects the fact that topology alone is not a very good predictor of term discriminating power, highlighting the need to recognize descriptive power and discriminating power as separate notions of term importance.

5.2 Evaluating the Distillation Method

In order to test the distillation method for query formation, we used again the Mars 2001 knowledge model. For each map, a baseline static method and three different dynamic feature selection methods were applied to select query terms. We use *Inverse Map Frequency* (IMF) as the baseline static feature selection method. IMF is an adaptation of the IDF weighting scheme [Salton and Yang, 1973], designed to measure the overall rarity of a term in a knowledge model. Each term t in a map was weighted as $IMF(t) = \log \frac{1+|\mathcal{K}|}{|\mathcal{K}_t|}$, where $|\mathcal{K}|$ represents the number of concept maps in the knowledge model (118 for $\mathcal{K}$ = "Mars 2001") and $|\mathcal{K}_t|$ stands for the number of concept maps containing term t. IMF was used to sort the terms occurring in a concept map and to generate queries of incremental size, starting from a query of size 1 consisting of the most highly weighted term and incrementally adding the next most highly weighted terms.

The dynamic weighting schemes evaluated here are three variations on the framework for query distillation proposed in this paper. We refer to these methods as *Dynamic Basic* (DB), *Dynamic Concept-Root* (DCR), and *Dynamic Concept-Root-Disjunction* (DCRD). All three methods are based on the algorithm discussed in section 4, but differ on how the queries are constructed for each concept in a concept map. Consider a concept map with concept root whose label consists of terms $r_1, r_2, \ldots, r_x$. Given a concept c with terms $t_1, t_2, \ldots, t_y$ the three types of queries associated with c are the following:

DB: t_1 AND t_2 AND $\ldots$ AND t_y.

DCR: t_1 AND t_2 AND $\ldots$ AND t_y AND r_1 AND r_2 AND $\ldots$ AND r_x.

DCRD: (t_1 AND t_2 AND $\ldots$ AND t_y AND r_1 AND r_2 AND $\ldots$ AND r_x) OR t_1 OR t_2 OR $\ldots$ OR t_y OR r_1 OR r_2 OR $\ldots$ OR r_x.

Because GOOGLE limits queries to 10 words, we truncated those queries that resulted in more than 10 term occurrences. In our evaluation we constructed a query for each concept in a concept map and considered up to 30 returned results per query. The search results associated with a concept were divided into 3 sets of equal size. In a three-stage evaluation, we used one of the three sets for query distillation and the other two for testing, rotating the roles of the sets at each stage. For each stage, the distillation data was used to compute an approximation of the discriminating power Δ of each term. Only the information readily available from the search results (snippets, etc.) was used in the distillation phase. The query involving terms with highest Δ value was identified as the *most promising query*, as done in the algorithm of figure 5. To test the query distillation method we selected from the testing data the remaining two sets of returned results (i.e., the search results not used for query distillation) associated with the most promising query and used those sets for performance analysis of the corresponding dynamic method.

To evaluate the performance of our methods, we took the full documents associated with the returned results, and computed their mean similarity to the source concept map. Similarity was measured as the proportion of novel terms (terms not in the query) in a retrieved document that are also part of the source map. Given a set **Q** of terms in a query, a set **M** of terms in a source map, and a set **D** containing the terms of a query result, the similarity of the query result to the source map can be measured by:

$$\mathcal{S}(\mathbf{Q}, \mathbf{M}, \mathbf{D}) = \frac{|(\mathbf{D} \cap \mathbf{M}) - \mathbf{Q}|}{|(\mathbf{D} \cup \mathbf{M}) - \mathbf{Q}|}.$$

Measure $\mathcal{S}$ is an adaptation of the *Jaccard coefficient*. It computes the proportion of terms in the source map or in a retrieved result that are in both the map and the retrieved result but are not in the query. If the set of search results for a given query is empty, the value for that query is considered to be 0.

In order to control for query size when comparing the performance of the dynamic methods against IMF, we set the size of the IMF queries to the number of terms occurring in the conjunctive portion of the corresponding dynamic-method query.

Figure 6 compares performance of the three dynamic methods to the IMF method. Each concept map in the Mars 2001 project corresponds to a trial and is represented by a point. The point's horizontal coordinate corresponds to the average performance of IMF for that case, while the vertical coordinate corresponds to the average performance of the dynamic method. In this evaluation DB outperforms IMF in 74% of the cases, DCR outperforms IMF in 77% of the cases, and DCRD outperforms IMF in 64% of the cases. In particular, there are several cases in which queries formed using the IMF method resulted in no search results. This highlights one of

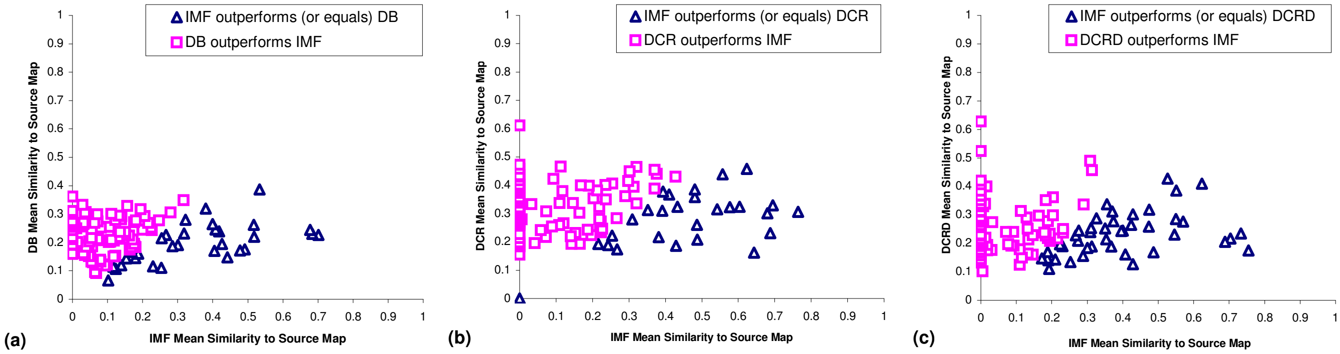


Figure 6: Average similarity to source map of documents retrieved using IMF vs. (a) DB, (b) DCR, and (c) DCRD.

the main advantages of using a dynamic approach involving a distillation phase to discover which are the most useful terms to use in a query. In Tables 2, 3 and 4 we present the mean similarity confidence interval resulting from each of the dynamic methods, and we compare it against the mean similarity confidence interval resulting from applying the IMF method with query size adjusted as we explained above. These comparison tables show that the three dynamic methods result in statistically significant improvements over IMF.

	N	MEAN	STDEV	SE	95% C.I.
DB	118	0.2196	0.0645	0.0059	**(0.2079, 0.2311)**
IMF	118	0.1627	0.1563	0.0144	**(0.1345, 0.1909)**

Table 2: DB vs. IMF: confidence intervals for the mean similarity to source map.

	N	MEAN	STDEV	SE	95% C.I.
DCR	118	0.3111	0.0893	0.0082	**(0.2950, 0.3272)**
IMF	118	0.1798	0.2037	0.0188	**(0.1430, 0.2165)**

Table 3: DCR vs. IMF: confidence intervals for the mean similarity to the source map.

	N	MEAN	STDEV	SE	95% C.I.
DCRD	118	0.2498	0.0903	0.0083	**(0.2335, 0.2661)**
IMF	118	0.1880	0.1955	0.0180	**(0.1527, 0.2232)**

Table 4: DCRD vs. IMF: confidence intervals for the mean similarity to the source map.

The fact that the dynamic methods rely on the submission of a first round of queries (distillation phase) to approximate a term's descriptive and discriminating power suggests that they are less efficient than the static approaches. However, given that knowledge will be extended incrementally, multiple rounds of queries will be submitted in any case, and the generation of second-round and subsequent queries can significantly benefit from examining previous search results, at a small additional cost.

6. RELATED WORK

Extensions to basic IR approaches have examined some of the issues raised in this paper. For instance, some automatic relevance feedback techniques, such as the Rocchio's method [Rocchio, 1971], make use of the full search context for query refinement. In these approaches the original query is expanded by adding a weighted sum of terms corresponding to relevant documents, and subtracting a weighted sum of terms from irrelevant documents. As a consequence the terms that occur often in documents similar to the input topic will be assigned the highest rank, as in our descriptors. However, our technique also gives priority to terms that *occur only in relevant documents* and not just to those that *occur often*. In other words, we prioritize terms for both discriminating and descriptive power. The techniques for query term selection proposed in this paper share insights and motivations with other methods for query expansion and refinement [Scholer and Williams, 2002, Billerbeck et al., 2003]. However, systems applying these methods differ from EXTENDER in that they support this process through a query or browsing interface requiring explicit user intervention, rather than formulating queries automatically.

Our techniques rely on the notions of term co-occurrence and document similarity to discover higher-order relationships in collections of documents. This relates to the use of LSA [Deerwester et al., 1990] to uncover the latent relationships between words in a collection. However, LSA's goal is to compute a matrix representing semantic distance between terms and documents, without identifying topic descriptors and discriminators.

The CmapTools search enhancer [Carvalho et al., 2001] uses concept maps to provide search context, but differs from EXTENDER in requiring user-generated queries and returning Web pages instead of topics. Systems that examine the user's current document to proactively provide suggestions include the Remembrance Agent [Rhodes and Starner, 1996] and Watson [Budzik and Hammond, 1999]. Other tools monitor user browsing activity to identify relevant Web pages (e.g., [Armstrong et al., 1995, Lieberman, 1995]). All these systems are similar to EXTENDER in attempting to provide users with context-relevant information, but differ in not attempting to generate new topics.

Topic-driven crawlers, also called focused crawlers [Chakrabarti et al., 1999, Menczer et al., 2004], follow hyperlinks to find information relevant to a triggering topic. EXTENDER contrasts in relying entirely on a search engine to mine the Web for topics—it does not crawl the Web—and in not being aimed at generating extensive topic information. Instead, it attempts to dynamically generate samples of topics that will serve as hints to the knowledge modeler.

7. CONCLUSION

This paper develops a framework for the extraction of topic descriptors and discriminators to aid information search in the context of a knowledge model under construction. It proposes and evaluates methods that apply this framework and discusses the use of these methods in the implementation of EXTENDER, CmapTools' topic suggester. The EXTENDER system proactively aids the user's knowledge extension process by providing novel but related topics which the user may not have considered.

EXTENDER's operation relies on the dynamic assessment of term descriptive and discriminating power to refine queries and to filter irrelevant material. During EXTENDER's first cycle, a term's descriptive power is obtained directly from the topology of the source concept map. However, for subsequent iterations, when topics are compiled as topology-free bags of terms, extracting good topic descriptors is important. When the system presents the final generation of topics to the user, the topic descriptors are used to produce labels for the suggested topics.

The evaluation presented in this paper took a bottom-up approach, focusing on the ability of EXTENDER to find good topic descriptors and discriminators at each step of its process. We consider this type of evaluation to be an important step for guiding the development of knowledge extension support tools, and EXTENDER appears to give good results in practice. We are now designing experiments to directly test human subjects' assessment of the relevance and usefulness of EXTENDER's suggested topics during the knowledge model extension process.

Acknowledgments

We gratefully acknowledge the contributions of our collaborators Alberto Cañas and the IHMC CmapTools team. This research is supported in part by NASA under award No NCC 2-1216. We thank Google for the use of their Web APIs service.

8. REFERENCES

Armstrong, R., Freitag, D., Joachims, T., and Mitchell, T. (1995). Webwatcher: A learning apprentice for the World Wide Web. In *AAAI Spring Symposium on Information Gathering*, pages 6–12.

Belkin, N. J. (2000). Helping people find what they don't know. *Commun. ACM*, 43(8):58–61.

Berge, C. (1973). *Graphs and Hypergraphs*. North Holland.

Billerbeck, B., Scholer, F., Williams, H. E., and Zobel, J. (2003). Query expansion using associated queries. In *CIKM*, pages 2–9.

Budzik, J. and Hammond, K. (1999). Watson: Anticipating and contextualizing information needs. In *62nd Annual Meeting of the American Society for Information Science*.

Cañas, A., Ford, K., Brennan, J., Reichherzer, T., and Hayes, P. (1995). Knowledge construction and sharing in Quorum. In *AIED*, pages 218–225.

Cañas, A., Leake, D., and Maguitman, A. (2001). Combining concept mapping with CBR: Experience-based support for knowledge modeling. In *FLAIRS*, pages 286–290.

Cañas, A. J., Coffey, J., Reichherzer, T., Hill, G., Suri, N., Carff, R., Mitrovich, T., and Eberle, D. (1998). El-tech: A performance support system with embedded training for electronics technicians. In *FLAIRS*, pages 79–83.

Carvalho, M., Hewett, R., and Cañas, A. (2001). Enhancing Web searches from concept map-based knowledge models. In *Proceedings of the SCI Conference*.

Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11–16):1623–1640.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Ford, K. M., Coffey, J. W., Cañas, A. J., Andrews, E. J., and Turner, C. W. (1996). Diagnosis and explanation by a nuclear cardiology expert system. *International Journal of Expert Systems*, 9:499–506.

Hoffman, R. R., Coffey, J. W., Ford, K. M., , and Carnot, M. J. (2001). Storm-LK: A human-centered knowledge model for weather forecasting. In *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society*.

Jones, S. K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Kobayashi, M. and Takeda, K. (2000). Information retrieval on the Web. *ACM Comput. Surv.*, 32(2):144–173.

Leake, D., Maguitman, A., and Reichherzer, T. (2003a). Topic extraction and extension to support concept mapping. In *FLAIRS*, pages 325–329.

Leake, D., Maguitman, A., and Reichherzer, T. (2004). Understanding knowledge models: Modeling assessment of concept importance in concept maps. In *CogSci2004*, pages 785–800.

Leake, D., Maguitman, A., Reichherzer, T., Cañas, A., Carvalho, M., Arguedas, M., Brenes, S., and Eskridge, T. (2003b). Aiding knowledge capture by searching for extensions of knowledge models. In *KCAP*, pages 44-53.

Lieberman, H. (1995). Letizia: An agent that assists Web browsing. In *IJCAI*, pages 924–929.

Menczer, F., Pant, G., and Srinivasan, P. (2004). Topic-driven crawlers: Machine learning issues. *ACM TOIT (To appear)*.

Novak, J. and Gowin, D. B. (1984). *Learning How to Learn*. Cambridge University Press.

Rhodes, B. and Starner, T. (1996). The remembrance agent: A continuously running automated information retrieval system. In *PAAM*, pages 487–495.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In Salton, G., editor, *The SMART retrieval system - experiments in automatic document processing*, pages 313–323. Prentice-Hall.

Salton, G. and Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Salton, G. and Yang, C. (1973). On the specification of term values in automatic indexing. *Journal of Documentation*, 29:351–372.

Scholer, F. and Williams, H. E. (2002). Query association for effective retrieval. In *CIKM*, pages 324–331.